Outliers

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised April 7, 2016

These notes draw heavily from several sources, including Fox's Regression Diagnostics; Pindyck and Rubinfeld; Statistics for Social Data Analysis, by George Bohrnstedt and David Knoke, 1982; Norusis's SPSS 11 chapter 22 on "Analyzing residuals;" Hamilton's chapter on "Robust regression." Also some of the text is either copied verbatim or adapted from the Stata 12 manual. I'm hitting highlights here, but the readings include lots of other good suggestions and details.

Description of the problem. One problem with least squares occurs when there are one or more large deviations, i.e. cases whose values differ substantially from the other observations. These points are called *outliers*. You should be worried about outliers because (a) extreme values of observed variables can distort estimates of regression coefficients, (b) they may reflect coding errors in the data, e.g. the decimal point is misplaced; or you have failed to declare some values as missing (c) they may be a result of model misspecification – variables have been omitted that would account for the outlier; or, the outlier belongs to a different population than the one you want to study.

Detecting Outliers using Stata

As is often the case with Stata, instead of a few big commands with several options, we execute several smaller commands instead. How useful different approaches are may depend, in part, on whether you are analyzing a few dozen cases, or several thousand. We'll take a closer look at the data used in the following regression:

```
. use https://www3.nd.edu/~rwilliam/statafiles/outliers.dta, clear
. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   11.29
       Model |  3220.79618     1  3220.79618           Prob > F      =  0.0018
    Residual |  10844.1543    38  285.372482           R-squared     =  0.2290
-------------+------------------------------           Adj R-squared =  0.2087
       Total |  14064.9505    39  360.639757           Root MSE      =  16.893

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .6686151   .1990218     3.36   0.002     .2657166    1.071514
       _cons |   3.727621   2.772329     1.34   0.187    -1.884665    9.339907
------------------------------------------------------------------------------
```

If you knew the data very well, you might already see something that makes you suspicious. Given that you don't, here are some things to check out.

Descriptive statistics. It is always a good idea to start with descriptive statistics of your data. Besides the built-in command `summarize`, the user-written commands `fre` and `extremes` can be helpful here. (To save space I am only printing out a few of the frequencies.)

```
. * Basic descriptive stats
. fre dv iv, tabulate(3)

dv
------------------------------------------------------------------
                  |      Freq.    Percent      Valid       Cum.
------------------+-----------------------------------------------
Valid   -20.44946 |          1       2.50       2.50       2.50
        -19.62192 |          1       2.50       2.50       5.00
        -17.21676 |          1       2.50       2.50       7.50
        :         |          :          :          :          :
        16.84143  |          1       2.50       2.50      95.00
        19.36333  |          1       2.50       2.50      97.50
        99        |          1       2.50       2.50     100.00
        Total     |         40     100.00     100.00
------------------------------------------------------------------

iv
------------------------------------------------------------------
                  |      Freq.    Percent      Valid       Cum.
------------------+-----------------------------------------------
Valid   -35.74697 |          1       2.50       2.50       2.50
        -35.24309 |          1       2.50       2.50       5.00
        -21.73665 |          1       2.50       2.50       7.50
        :         |          :          :          :          :
        17.97275  |          1       2.50       2.50      95.00
        18.77257  |          1       2.50       2.50      97.50
        22.44931  |          1       2.50       2.50     100.00
        Total     |         40     100.00     100.00
------------------------------------------------------------------

. sum dv iv

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          dv |        40    1.232763    18.99052  -20.44946         99
          iv |        40   -3.731381    13.59168  -35.74697   22.44931
```

Both the frequencies and the summary statistics indicate that dv has a maximum value of 99, which is much higher than the other values of dv. No values immediately stick out for iv.

Nick Cox's `extremes` command provides perhaps an easier way of identifying the cases with the most extreme high and low values.

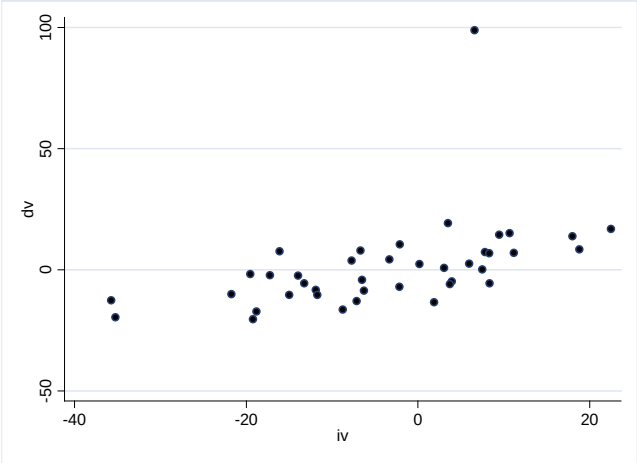
```
. extremes dv iv

  +------------------------------+
  | obs:          dv          iv |
  |------------------------------|
  |   5.   -20.44946   -19.22762 |
  |   7.   -19.62192   -35.24309 |
  |   8.   -17.21676   -18.83887 |
  |  38.   -16.34352   -8.757764 |
  |  16.   -13.33637    1.862242 |
  +------------------------------+
```

```
+----------------------------+
|  32.   14.51918   9.434002 |
|  36.   15.22688   10.65133 |
|  13.   16.84143   22.44931 |
|  22.   19.36333   3.484266 |
|***9.*********99***6.599043*|
+----------------------------+
```

Notice the format of the command and the layout of the output. You could just specify one variable, and it would give you the extreme values for it. If you specify two or more variables, it will give you the extreme values of the first variable, and the values of the other variables for those same cases. This can be useful for determining if the extreme values really are that extreme, given the values of the other variables. We see that case 9 seems very different from the rest of the cases and has a very suspicious value of 99. You can repeat the process for other variables in the analysis (in this case nothing obvious stands out for iv).

Graphic techniques. Particularly when the sample is small, graphic techniques can be helpful. First, we can use the `scatter` command to plot the dv and the iv.

```
. scatter dv iv
```

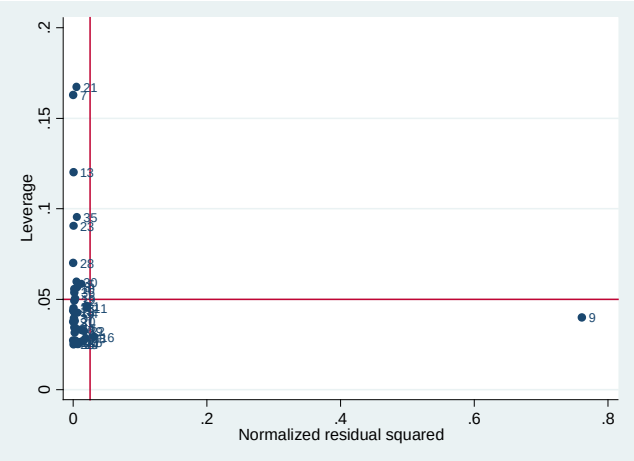


Note the outlying case in the upper right.

After we have run the regression, we have several post-estimation commands than can help us identify outliers. According to the Stata 12 Manual, "One of the most useful diagnostic graphs is provided by lvr2plot (leverage-versus-residual-squared plot), a graph of leverage against the (normalized) residuals squared." (The mlabel option made the graph messier, but by labeling the dots it is easier to see where the problems are.)

NOTE: Cases that are outliers on X will have more leverage than cases that have values close to the mean of X. Appendix A discusses leverage in more detail.

```
. gen id = _n
. lvr2plot, mlabel(id)
```



The Stata 12 manual says "The lines on the chart show the average values of leverage and the (normalized) residuals squared. Points above the horizontal line have higher-than-average leverage; points to the right of the vertical line have larger-than-average residuals."

The graph shows us that case 9 has a very large residual (i.e. the difference between the predicted and observed value for case 9 is exceptionally large) but it doesn't have much leverage. Cases in the upper right of the graph (if there were any) would be especially important because they would be high leverage and large residuals. If you see a group of outliers together you may wish to check to see what, if anything, they share in common.

Residual Statistics. We can also compute a variety of residual statistics. Basically, we use the `predict` command to compute the measures we want, and then run the summary statistics on them. I'll show a few examples; typing `help regress` will show you other options.

CAUTION: In general, `predict` calculates the requested statistic for all observations possible, whether they were used in fitting the model or not. This can be quite handy at times. But, if your regression was not run on all the cases, e.g. you were analyzing a subsample, you might want to modify the following commands to something like `predict stdresid if e(sample), rstandard`. The `if` parameter will limit the computations to the cases used by the previous regression.

```
. * Residual statistics
. * Discrepancy measures
. * Standardized residuals -- values more extreme than 3 may be a problem
. predict stdresid, rstandard
. * Studentized residual
. predict rstudent, rstudent

. * leverage measure
. * leverage (or hat) identifies cases that can have a large effect on the
. * fitted model even if the corresponding residual is small
. * When the leverage > 2k/n then there is high leverage
. * Maybe use 3k/n when N is small
. predict leverage, leverage
```

```
. * Influence measures
. * DFBetas -- SPSS calls these SDBETAS -- values larger than 1
. * or > 2/ sqrt(N) (about .316 in this case) are a problem
. . dfbeta
                      _dfbeta_1: dfbeta(iv)
. * Get Cook's Distance measure -- values greater than 4/N may cause concern
. predict cooksd, cooksd

. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          dv |        40    1.232763    18.99052  -20.44946         99
          iv |        40   -3.731381    13.59168  -35.74697   22.44931
          id |        40        20.5    11.69045          1         40
    stdresid |        40    .0007058      1.0076   -1.10009   5.488952
    rstudent |        40    .1617825    1.960709  -1.103228   11.90047
-------------+--------------------------------------------------------
   _dfbeta_1 |        40    .0190428    .2449757  -.1998377   1.478083
      cooksd |        40    .0208666      .09811   .0000153   .6246131
    leverage |        40         .05    .0336828   .0250226   .1672696
```

Why does Stata offer so many residual statistics??? Different statistics tell you different things about the outliers, and one statistic may catch problems that are missed by another.

- Some statistics measure *discrepancy*, i.e. the difference between the predicted Y and the observed Y.
- But, some outliers will have relatively little influence on the regression line. An extreme value of y that is paired with an average value of X will have less effect than an extreme value of Y that is paired with a non-average value of X. An observation with an extreme value on a predictor variable (or with extreme values on multiple Xs) is called a point with high *leverage*. Some residual statistics therefore measure leverage.
- Fox gives the useful formula Influence on Coefficients = Leverage x Discrepancy. By this he means that outlying values on Y will have the greatest impact when (a) their corresponding X values are further away from the mean of X, and (b) the Y value is out of line with the rest of the Y values, i.e. it does not fall on the same line that the other cases do. We will discuss this more later in the handout. Some residual statistics therefore measure *influence*.

To explain a few of the statistics presented by Stata:

Discrepancy Measures. According to the Stata 12 Manual, "Standardized and Studentized residuals are attempts to adjust residuals for their standard errors… In general, Studentized residuals are preferable to standardized residuals for purposes of outlier identification.

- Studentized residuals can be interpreted as the t statistic for testing the significance of a dummy variable equal to 1 in the observation in question and 0 elsewhere (Belsley, Kuh, and Welsch 1980). Such a dummy variable would effectively absorb the observation and so remove its influence in determining the other coefficients in the model." Values of 3 or greater (or -3 or less) may be problematic.

Leverage Measure. The leverage option (which can also be called hat) calculates the Hosmer and Lemeshow leverage or the diagonal element of the hat matrix (so named because its computation involves y_hat).

- Univariate or multivariate X outliers are high-leverage observations.
- Leverage is bounded by two limits: 1/n and 1. The closer the leverage is to unity, the more leverage the value has.
- When the leverage > 2k/n then there is high leverage. For small samples you may want to use 3k/n. Others say a point with leverage greater than (2k+2)/n should be carefully examined. Here **k** is the number of predictors and **n** is the number of observations. In this case we might be worried about cases with leverage values of 2/40 (.05) or 3/40 (.075) or 4/40 (.10).

Influence Measures. DFBETA shows how much a coefficient would change if that case were dropped from the data.

- According to the Stata 12 manual, "DFBETAs are perhaps the most direct influence measure of interest to model builders. DFBETAs focus on one coefficient and measure the difference between the regression coefficient when the ith observation is included and excluded, the difference being scaled by the estimated standard error of the coefficient. Belsley, Kuh, and Welsch (1980, 28) suggest observations with dfbetas > 2/Sqrt(N) should be checked as deserving special attention, but it is also common practice to use 1 (Bollen and Jackman 1990, 267), meaning that the observation shifted the estimate at least one standard error."
- In this example we would look for a dfbeta > .316 or else > 1.) Note that the larger the sample, the harder it is for any one case to affect the regression coefficients.

Cook's distance is another way of measuring influence. According to the Stata 12 Manual, "Cook's distance measures the aggregate change in the estimated coefficients when each observation is left out of the estimation. Values of Cook's distance that are greater than 4/N (in this case, 4/40 = .10) may be problematic.

The statistics show us that at least one standardized and studentized residual is much larger than 3, at least one of the dfbetas is larger than 2/sqrt(40) (which means that deletion of that case would cause a substantial change in the parameter estimates), and that at least one Cook's distance is much larger than 4/N = .10. Again using the `extremes` command, it is pretty obvious case 9 is our biggest problem.

```
. extremes stdresid rstudent _dfbeta_1 cooksd leverage
```

obs:	stdresid	rstudent	_dfbeta_1	cooksd	leverage
16.	-1.10009	-1.103228	-.0737939	.0182921	.0293429
11.	-.9054657	-.9032696	-.1315701	.0194316	.0452566
38.	-.8537651	-.8506546	.0511074	.0106943	.0285067
39.	-.7236397	-.719026	-.0640732	.0088459	.0326813
25.	-.7062296	-.701494	.028504	.0068167	.0266071

```
  +-------------------------------------------------------------+
  |  29.   .4976577    .492674    .0094423   .0032218    .025358 |
  |  40.    .519945   .5148928   -.0183424   .0036419   .0262358 |
  |  22.   .8006765   .7968211    .0688563    .010674   .0322267 |
  |   1.   .8894223   .8869218   -.1324265   .0191859   .0462622 |
  |***9.***5.488952***11.90047***1.478083***.6246131***.0398124*|
  +-------------------------------------------------------------+
```

NOTE: To be thorough, we should probably run the `extremes` command specifying the other residual measures first as well; the cases that are the most extreme on stdresid won't necessarily be the most extreme on other measures.

Dealing with outliers (Both Stata and SPSS)

- First, check to make sure there are no coding errors. Has an extra zero been added to the outlying case?
- Make sure missing data coding is correct. For example, if you have a variable whose coding runs from 0 to 7 with an MD code of 99, and you have failed to tell SPSS that 99 is an MD code (or have not recoded 99 to . in Stata), the regression estimates will be way, way off. I've seen this produce extremely high correlations, when both the IVs and DVs were not being properly treated as missing.
- Run the regression both with and without the outlying cases. If the results are substantially different, this should be noted. You should either explain why some cases were deleted, or present both sets of analyses. For example, in this case,

```
. use https://www3.nd.edu/~rwilliam/statafiles/outliers.dta, clear
. reg dv iv if dv!=99

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  1,    37) =   33.19
       Model |  2015.14589     1  2015.14589           Prob > F      =  0.0000
    Residual |  2246.28396    37  60.7103773           R-squared     =  0.4729
-------------+------------------------------           Adj R-squared =  0.4586
       Total |  4261.42986    38  112.142891           Root MSE      =  7.7917

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5329324   .0925018     5.76   0.000     .3455059    .7203589
       _cons |   .8556491   1.301279     0.66   0.515    -1.780992     3.49229
------------------------------------------------------------------------------
```

Among other things, you see that both the slope and the intercept are smaller than in the original regression, and the R^2 value is higher.

- You might reasonably decide that the outlying case does not fall within the population of interest that you want to study, and can justify excluding it that way.
- Large outliers might be accounted for by adding more explanatory variables. Naturally, you prefer to explain the values of cases, rather than just discard them.
- Sometimes a transformation of a variable may be warranted, e.g. take the log of the variable.

- Remember, though, that outliers may represent important information about the relationship between variables. Don't throw the outlier away without examining it first. Maybe you will catch a coding error. Perhaps you can explain why this case doesn't really fall into the population of interest. Or, perhaps you can add IVs which will explain why this case's values differ so much from the rest.

Note: Appendix B discusses Robust Regression techniques which can be estimated using Stata. You should read through the appendix; but I am emphasizing these techniques somewhat less than I have in the past because they seem to be rarely used in Sociology and their utility has been questions (at least for the official routines that Stata has built in).

Appendix C has information on dealing with outliers in SPSS.

Appendix A: Importance of Leverage

As previously pointed out, Fox says that Influence on Coefficients = Leverage * Discrepancy. Among other things, this means that outliers on Y that are paired with average values of X will have less influence on parameter estimates than outliers on Y that are paired with above or below-average values on X. In the current example, the value of iv for case 9 is 6.599043; the average value of the other 39 cases is -3.996264. So, the iv value for case 9 is above average, but not extremely so (the highest value for iv is 22.44931). Let's see what happens to the regression estimates as we make case 9 more and more of an outlier on dv:

```
. use https://www3.nd.edu/~rwilliam/statafiles/outliers.dta, clear
. * Case 9: dv = 999
. replace dv = 999 in 9
(1 real change made)

. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    1.10
       Model |   27651.582     1   27651.582           Prob > F      =  0.3001
    Residual |  952144.395    38  25056.4315           R-squared     =  0.0282
-------------+------------------------------           Adj R-squared =  0.0026
       Total |  979795.977    39  25122.9738           Root MSE      =  158.29

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   1.959091   1.864894     1.05   0.300     -1.81619    5.734372
       _cons |   31.04288   25.97755     1.19   0.239    -21.54593    83.63169
------------------------------------------------------------------------------

. * Case 9: dv = 9999
. replace dv = 9999 in 9
(1 real change made)

. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    0.63
       Model |  1591744.43     1  1591744.43           Prob > F      =  0.4321
    Residual |  95917861.8    38  2524154.26           R-squared     =  0.0163
-------------+------------------------------           Adj R-squared = -0.0096
       Total |  97509606.2    39  2500246.31           Root MSE      =  1588.8

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   14.86385    18.7177     0.79   0.432    -23.02816    52.75586
       _cons |   304.1954   260.7334     1.17   0.251    -223.6316    832.0225
------------------------------------------------------------------------------
```

As we see, the more extreme the outlier is, the more the regression estimates are affected.

Let's now see what happens when we change case 9 so that it has the mean value on iv, and we make it more and more of an outlier on iv:

```
. * Make case 9 exactly average on iv
. replace iv = -3.996264 in 9
```

```
(1 real change made)

. * Make dv = 99
. replace dv = 99 in 9
(1 real change made)

. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    6.35
       Model |  2015.14588     1  2015.14588           Prob > F      =  0.0160
    Residual |  12049.8046    38  317.100122           R-squared     =  0.1433
-------------+------------------------------           Adj R-squared =  0.1207
       Total |  14064.9505    39  360.639757           Root MSE      =  17.807

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5329324   .2114059     2.52   0.016     .1049635    .9609012
       _cons |   3.362501     2.9396     1.14   0.260    -2.588407     9.31341
------------------------------------------------------------------------------

. * Make dv = 999
. replace dv = 999 in 9
(1 real change made)

. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    0.08
       Model |  2015.14579     1  2015.14579           Prob > F      =  0.7811
    Residual |  977780.831    38  25731.0745           R-squared     =  0.0021
-------------+------------------------------           Adj R-squared = -0.0242
       Total |  979795.977    39  25122.9738           Root MSE      =  160.41

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5329324   1.904355     0.28   0.781    -3.322232    4.388097
       _cons |    25.8625   26.48006     0.98   0.335    -27.74358    79.46858
------------------------------------------------------------------------------

. * Make dv = 9999
. replace dv = 9999 in 9
(1 real change made)

. reg dv iv

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =    0.00
       Model |  2015.14489     1  2015.14489           Prob > F      =  0.9778
    Residual |  97507591.1    38  2565989.24           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0263
       Total |  97509606.2    39  2500246.31           Root MSE      =  1601.9

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5329322   19.01719     0.03   0.978    -37.96535    39.03121
       _cons |   250.8625   264.4341     0.95   0.349    -284.4563    786.1813
------------------------------------------------------------------------------
```

As you see, the slope coefficient barely changes at all, although the intercept, t-values, and various other statistics do. Since, in these examples, case 9 has an average value on iv, it has no leverage and hence virtually no effect on the slope estimate.

If you are curious as to why this is – recall that the formula for the bivariate slope coefficient is

$$\beta_{yx} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})}$$

So if, for a specific case, x_i = mean of x, that case adds 0 to both the numerator and the denominator of the above formula – no matter what y_i equals for that case. Hence, you can plug in whatever values you want for y for that case (or bigger and bigger values, like I did) and it will have no effect on the slope coefficient. Conversely, the more x_i differs from the mean of x, the more impact that case can have on the slope coefficient.

Appendix B: Dealing with outliers (Stata) – Robust Regression Techniques

One advantage of Stata over SPSS is that it includes so-called robust regression routines that are better able to handle outliers. The built-in routines are `rreg` and `qreg`, although (as noted below) many argue that there are user-written routines that are better. (We would, of course, still want to do all the things described above, but if the outliers do appear to be legitimate, these techniques can help.) The `rreg` and `qreg` routines work best when it is the DV that has outliers rather than the IVs. As Hamilton notes (Statistics With Stata, Version 8, p. 239):

> OLS tends to track outliers, fitting them at the expense of the rest of the sample. Over the long run, this leads to greater sample-to-sample variation or inefficiency when samples often contain outliers. Robust regression methods aim to achieve almost the efficiency of OLS with ideal data and substantially better than OLS efficiency in non-ideal (for example, nonnormal errors) situations….[The Stata routines] `rreg` and `qreg` resist the pull of outliers, giving them better than OLS efficiency in the face of nonnormal, heavy-tailed error distributions.

To show how this works, first, let us repeat our regression results, this time excluding the outlying case:

```
. use https://www3.nd.edu/~rwilliam/statafiles/outliers.dta, clear
. reg dv iv if dv!=99

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  1,    37) =   33.19
       Model |  2015.14589     1  2015.14589           Prob > F      =  0.0000
    Residual |  2246.28396    37  60.7103773           R-squared     =  0.4729
-------------+------------------------------           Adj R-squared =  0.4586
       Total |  4261.42986    38  112.142891           Root MSE      =  7.7917

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5329324   .0925018     5.76   0.000     .3455059    .7203589
       _cons |   .8556491   1.301279     0.66   0.515    -1.780992     3.49229
------------------------------------------------------------------------------
```

Now, we'll see what happens when we run Stata's `rreg` (robust regression) routine with all 40 cases:

```
. rreg dv iv, nolog

Robust regression estimates                           Number of obs =      40
                                                      F(  1,    38) =   28.83
                                                      Prob > F      =  0.0000

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5352655   .0996925     5.37   0.000     .3334486    .7370824
       _cons |   .7893176   1.388694     0.57   0.573    -2.021946    3.600581
------------------------------------------------------------------------------
```

Note that we get estimates that are very similar to what we got when we used the regress command and dropped case 9.

`rreg` is a bit hard to explain. Basically, it goes through an iterative procedure (as Hamilton notes, it uses "iteratively reweighted least squares with Huber and biweight functions tuned for 95% Gaussian efficiency"), where the more extreme an outlier is, the less heavily it gets weighted in the regression calculations. Very extreme cases get dropped altogether. In this problem, `rreg` basically dropped case 9 altogether, which is why its final results looked so similar to the results we got when we ran a regression with case 9 excluded.

Another alternative is `qreg`, which stands for quantile regression (you'll also hear it referred to as Least Absolute Value Models or minimum L1-norm models). The most common form of quantile regression is median regression, where the goal is to estimate the median (rather than the mean) of the dependent variable, conditional on the values of the independent variables. Put another way, median regression finds a line through the data that minimizes the sum of the absolute residuals rather than the sum of the squares of the residuals as in ordinary regression (hence the term Least Absolute Value as opposed to Least Squares) . Medians are less affected by outliers than means are, so `qreg` can do better than `regress` when there are extreme outliers.

```
. qreg dv iv, nolog

Median regression                                   Number of obs =        40
  Raw sum of deviations 444.7982 (about -2.2611923)
  Min sum of deviations 335.8461                    Pseudo R2     =    0.2449

------------------------------------------------------------------------------
          dv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .6079845   .0891699     6.82   0.000     .4274695    .7884994
       _cons |   1.805331   1.326887     1.36   0.182    -.8808107    4.491472
------------------------------------------------------------------------------
```

As to which routine is better, and when? Well, that is a good question. In this particular case, `rreg` seems better, but all it basically did was drop the extreme case, which we could have done by ourselves. Had 99 been a legitimate code, `qreg` might have seemed the more appropriate choice. Hamilton also argues that when the IVs have outliers, `rreg` tends to do better because it tends to just drop such observations rather than try to fit them.

Again, I would stress that either of these routines should only be used after you have checked out other issues, e.g. are there coding errors in the data, is missing data being handled properly, would the addition of some other variable to the model make the outliers not be outliers anymore?

Advanced Discussion. But, if none of these solve the problem, here is some advice that was offered on Statalist when I asked about this on Jan. 30, 2004.

Nicholas Cox (he wrote a couple of messages and hopefully I have combined them correctly): This raises the old classical trope, beaten almost to death by the late Sir Isaiah Berlin in many of his essays on intellectual history, that the fox knows many things, but the hedgehog knows one big thing.

When attacked, the hedgehog has just one means of defence, although it is usually effective. -qreg- is a hedgehog. The fox has many different tricks. -rreg- is a fox. Its mixed strategy is an attempt to be smart in different ways.

My experience loosely matches Richard's, certainly in terms of wanting to think that -qreg- is as good because of the much greater ease in explaining it. At the same time, [if you have] "well-behaved" data + a "few" outliers (n ~ 1) it is sensible to use robust regression as a check on standard [regression]. [But if you have long-tailed data] you are possibly working on inappropriate scales and should wonder about reaching for a transformation or, in some frameworks, a different link function.

Also from Cox, as to whether outliers are more likely to be "real" or just coding mistakes: I think it depends, partly, on the kind of data you deal with. In fact these tribal differences among groups of statistical users are one of the persistently interesting features of Statalist.

In geography (that's my field) the big cities, countries, rivers, storms, etc. really are big and they really are important, and my advice to students and colleagues hinges on the idea that most outliers are likely to be genuine and important. Often this means taking logarithms! Also, there's usually a story behind each outlier and extra information somewhere…

In some other fields it may be that most outliers are mistakes and/or that in terms of advancing science it's better strategy to ignore them. The person who reports watching 180 hours of television a week is likely to be confused about something or other; and short of re-interviewing or some smart way of finding out that they really meant 18.0 or 108, the only possible thing may be to omit that data point.

Michael Blasnik: One difference between qreg and rreg is that they attempt to estimate different versions of the central tendency -- qreg estimates the median while rreg comes closer (in theory) to estimating a robust mean. The difference may be negligible in essentially symmetrical distributions, but for skewed distributions where the mean and median are not expected to be equal, one would expect their estimates to deviate systematically. If you really want to model the mean but are concerned about outliers, then rreg may be a better choice than qreg. If you want to model the median (or think the underlying distribution is fairly symmetrical), then qreg may be preferred.

I usually look at both and then try to figure out any substantive differences in results, but I'm generally partial to the coefficient estimates from rreg (I often deal with skewed distributions where the median is noticeably lower than the mean). On the other hand, I sometimes find rreg's std errors estimates questionable.

However, in a more recent exchange on January 14, 2011, `rreg` received considerable criticism.

Nicholas Cox: I'd advise against basing anything much on -rreg-.

The help file has it right: -rreg- is "one version of robust regression". When -rreg- was written the method seemed a good all-round flavour of robust regression, but it is doubtful whether it now looks like _the_ method of choice to anyone in 2011. If you ever used -rreg- for real, you'd be obliged to explain it and defend the choice in any serious forum. "I used robust regression" means virtually nothing. There are probably hundreds of ways to do robust regression (quite apart from what robustness means). "I used -rreg- as implemented in Stata" counts for little outside this community. "I used robust regression as codified by Li (1985)" obliges you to explain why you didn't use something more recent (to fad- and fashion-followers) or something else that someone else fancies for some reason of their own. The literature would keep you busy indefinitely. Outliers could be handled in many different ways. Considering transformations on one or

more variables is another way to do that. Wonder whether a linear structure makes sense scientifically is yet another.

On August 22, 2010, Steve Samuels suggested some other alternatives. I can't vouch for them because I haven't used them, but they may be worth investigating further if you feel that outliers are a problem and no other solution seems adequate.

> There are few rules about outliers, but the most important one is: OLS is the worst way to detect them. Detection requires a robust regression program; and a good program will not "reject" all outliers, but will automatically downweight them. For covariates, one wants to identify not outliers per se, but those with high leverage. But the decision about what to do with these is not automatic; sometimes they are the most important points and _must_ be kept.
>
> See: "Robust regression in Stata" by Vincenzo Verardi and Christophe Croux, The Stata Journal Volume 9 Number 3: pp. 439-453. Also available at:
> https://lirias.kuleuven.be/bitstream/123456789/202142/1/KBI_0823.pdf
>
> See also Verardi and Croux's contributed programs -mmregress- (findit) and Ben Jann's -robreg- (findit). These are superior to Stata's long-time built-in command -rreg-.

Appendix C: Using SPSS with Outliers [Optional]

SPSS also has some good routines for detecting outliers.

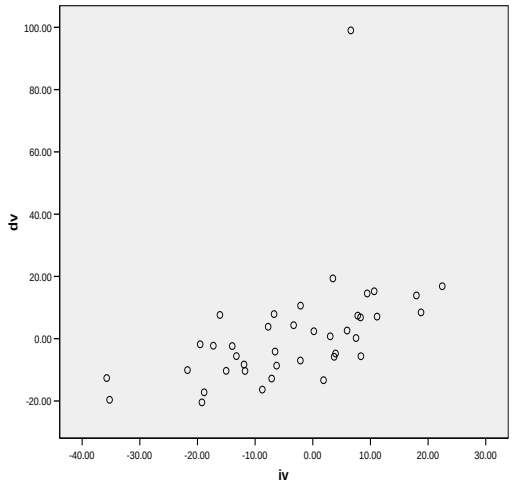
- There is always the FREQUENCIES routine, of course.
- The GRAPH command can do scatterplots of 2 variables.
- The EXAMINE procedure includes an option for printing out the cases with the 5 lowest and 5 highest values.
- The REGRESSION command can print out scatterplots (particularly good is *ZRESID by *ZPRED, which is a plot of the standardized residuals by the standardized predicted values). In addition, the regression procedure will produce output on CASEWISE DIAGNOSTICS, which indicate which cases are extreme outliers and/or which cases have the most impact on the regression estimates. This is particularly useful in that you see which cases stand out even after all IVs have been controlled for.

SPSS Example. Following is a hypothetical example of 40 cases. I constructed the data set so the DV and IV would have a correlation of about .7. I then changed one of the DV values into an extreme outlier. Note how the first three analyses (GRAPH, EXAMINE, and REGRESSION) all provide means of detecting the outlier. Then, see how the results change once the outlier is deleted and the regression is rerun.

```
Get File = 'D:\Soc63993\Spssfiles\outliers.sav'.

* This program shows some of the ways SPSS can be used to identify outliers.
* Do a scatterplot of vars to visually ID cases.
* Note that one case is way out of line with the rest.

GRAPH   /SCATTERPLOT(BIVAR)=iv WITH dv /MISSING=LISTWISE .
```

Graph

```
D:\Soc63993\Spssfiles\Outlier.sav
```



* Use examine procedure to id cases with extreme values on X or Y.
* However, note that these need not be outliers on a regression line.
* Note that Case 9 has a very extreme, and also very suspicious, value for DV.

```
EXAMINE
  VARIABLES=dv iv
  /PLOT NONE
  /STATISTICS EXTREME
  /MISSING LISTWISE
  /NOTOTAL.
```

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DV	40	100.0%	0	.0%	40	100.0%
IV	40	100.0%	0	.0%	40	100.0%

Extreme Values

			Case Number	Value
DV	Highest	1	9	99.00
		2	22	19.36
		3	13	16.84
		4	36	15.23
		5	32	14.52
	Lowest	1	5	-20.45
		2	7	-19.62
		3	8	-17.22
		4	38	-16.34
		5	16	-13.34
IV	Highest	1	13	22.45
		2	35	18.77
		3	23	17.97
		4	10	11.15
		5	36	10.65
	Lowest	1	21	-35.75
		2	7	-35.24
		3	28	-21.74
		4	30	-19.53
		5	5	-19.23

* Run regression with outlier in.
* Outlier will also show up in the plot.
* /Casewise prints out stats that help to ID extreme outliers, if any.
* /Scatterplot graphically helps to ID extreme outliers.
* SPSS Regression has many other options for analyzing residuals
* that may sometimes be useful.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT dv
  /METHOD=ENTER iv
  /Casewise defaults dfbeta
  /SCATTERPLOT=(*ZRESID ,*ZPRED ) .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
DV	1.2328	18.99052	40
IV	-3.7314	13.59168	40

Correlations

		DV	IV
Pearson Correlation	DV	1.000	.479
	IV	.479	1.000
Sig. (1-tailed)	DV	.	.001
	IV	.001	.
N	DV	40	40
	IV	40	40

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	IV[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: DV

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479[a]	.229	.209	16.89297

a. Predictors: (Constant), IV

b. Dependent Variable: DV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3220.796	1	3220.796	11.286	.002[a]
	Residual	10844.154	38	285.372		
	Total	14064.950	39			

a. Predictors: (Constant), IV

b. Dependent Variable: DV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.728	2.772		1.345	.187
	IV	.669	.199	.479	3.360	.002

a. Dependent Variable: DV

Casewise Diagnostics[a]

Case Number	Std. Residual	DV	Predicted Value	Residual	DFBETA	
					(Constant)	IV
9	5.379	99.00	8.1398	90.8602	2.872	.136

a. Dependent Variable: DV

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-20.1733	18.7376	1.2328	9.08760	40
Std. Predicted Value	-2.356	1.926	.000	1.000	40
Standard Error of Predicted Value	2.67222	6.90899	3.63283	1.04815	40
Adjusted Predicted Value	-21.6857	18.9965	1.2077	9.22506	40
Residual	-18.3091	90.8602	.0000	16.67499	40
Std. Residual	-1.084	5.379	.000	.987	40
Stud. Residual	-1.100	5.489	.001	1.008	40
Deleted Residual	-18.8626	94.6275	.0251	17.37645	40
Stud. Deleted Residual	-1.103	11.900	.162	1.961	40
Mahal. Distance	.001	5.549	.975	1.314	40
Cook's Distance	.000	.625	.021	.098	40
Centered Leverage Value	.000	.142	.025	.034	40

a. Dependent Variable: DV

To explain a few of the statistics presented by SPSS:

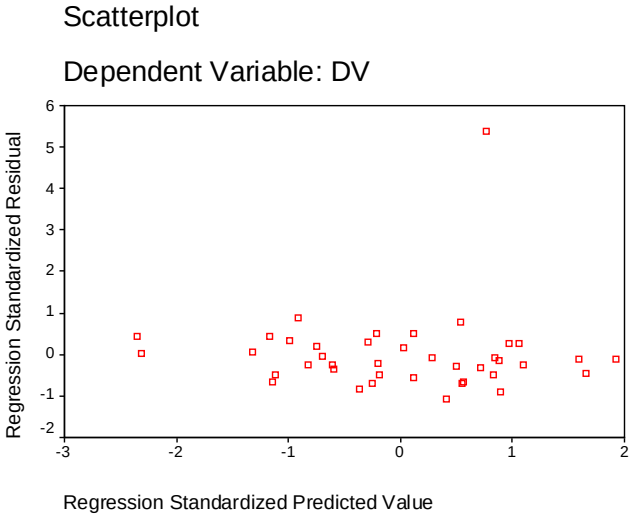
Std residual and Stud Residual are slightly different ways of standardizing the residuals; values of 3 or greater (or -3 or less) may be problematic. These items measure discrepancy but not necessarily how much influence the outlier has on the regression estimates.

DFBETA shows how much a coefficient would change if that case were dropped from the data. In this case, it shows that the effect of IV would drop by .136 if case 9 were dropped. [CAUTION: To make things confusing, Stata uses the term dfbeta to refer to what SPSS would call standardized dfbetas. There are other instances where Stata and SPSS use different naming conventions. With a standardized dfbeta, values of 1 or larger are generally considered important; others recommend that standardized dfbetas > 2/Sqrt(N) should be checked. The latter may be more reasonable, since the larger the sample, the harder it is for any one case to affect the coefficients, even if it is an extreme outlier.]

Cook's distance is another way of measuring influence. Values of Cook's distance that are greater than 4/N (in this case, 4/40 = .10) may be problematic.

From the above, we see that we have some very large standardized residuals and a large value for Cook's distance; further, we see that case 9 in particular is a problem (the fact that it was printed in the casewise diagnostics means it has a standardized residual of at least 3.)

Charts



* Get rid of the outlying case.

```
USE ALL.
COMPUTE filter_$=(dv < 99).
VARIABLE LABEL filter_$ 'dv < 99 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
```

* Rerun the regression without the outlier. Note changes in
* the correlation and in the coefficients. Now that the weird case is gone,
* the slope goes down because the regression line doesn't need to try to
* reach the outlier. Also note that "Casewise diagnostics"
* does not show up anymore because there are no extreme outliers left.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT dv
  /METHOD=ENTER iv
  /Casewise
 /SCATTERPLOT=(*ZRESID ,*ZPRED ) .
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
DV	-1.2741	10.58975	39
IV	-3.9963	13.66436	39

Correlations

		DV	IV
Pearson Correlation	DV	1.000	.688
	IV	.688	1.000
Sig. (1-tailed)	DV	.	.000
	IV	.000	.
N	DV	39	39
	IV	39	39

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	IV[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: DV

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.688[a]	.473	.459	7.79169

a. Predictors: (Constant), IV

b. Dependent Variable: DV

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2015.146	1	2015.146	33.193	.000[a]
	Residual	2246.284	37	60.710		
	Total	4261.430	38			

a. Predictors: (Constant), IV

b. Dependent Variable: DV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.856	1.301		.658	.515
	IV	.533	.093	.688	5.761	.000

a. Dependent Variable: DV

Residuals Statistics[a]

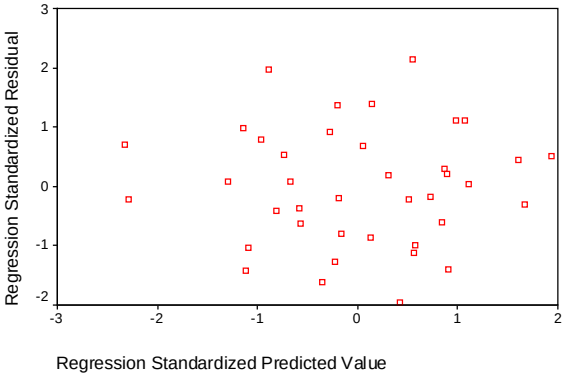
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-18.1951	12.8196	-1.2741	7.28218	39
Std. Predicted Value	-2.324	1.935	.000	1.000	39
Standard Error of Predicted Value	1.24920	3.19102	1.69670	.49064	39
Adjusted Predicted Value	-19.3137	12.2492	-1.3109	7.30290	39
Residual	-15.1845	16.6508	.0000	7.68848	39
Std. Residual	-1.949	2.137	.000	.987	39
Stud. Residual	-1.979	2.174	.002	1.008	39
Deleted Residual	-15.6618	17.2284	.0368	8.02534	39
Stud. Deleted Residual	-2.065	2.296	.004	1.029	39
Mahal. Distance	.002	5.399	.974	1.306	39
Cook's Distance	.000	.099	.022	.025	39
Centered Leverage Value	.000	.142	.026	.034	39

a. Dependent Variable: DV

Charts

Scatterplot

Dependent Variable: DV



Having dropped the problematic case, we see that all is pretty much well now.